Unique identifiers for training materials

Elin Kronander

title: "Releases and use of unique identifiers"

author:

name: "Elin Kronander"

orcid: 0000-0003-0280-6318

email: elin.kronander@nbis.se

Part of FAIR Training Material by Design workshop 2024-09-18

- https://zenodo.org/doi/10.5281/zenodo.13773159

Learning outcomes

By the end of this session, you should be able to:

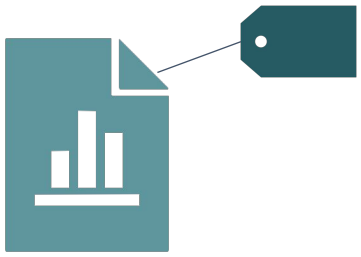
- **Explain** what unique persistent identifiers are and their benefits
- **List** and differentiate the types of unique identifiers that are relevant for publishing and sharing training materials
- **Compare** different strategies for unique identifiers for training materials
- **Create** versioned DOIs for training materials

NBIS

This session will be divided in 2 parts:

- Part 1, will be more theoretical and cover what unique identifiers are and how they can be used for training materials. We will see some real world examples and discuss different strategies.
  - What is a unique identifier?
  - Why are they useful for training materials?
  - Explore different strategies for adding them to training materials
- Part 2, will be more practical where you will use the public repository Zenodo to assign a digital object identifier to your github repo or if you have used Google drive there will be an option for that as well
  - Use Zenodo to assign a DOI to training materials
  - Use releases in GitHub to make versioned DOIs

What is a unique persistent identifier (PID)



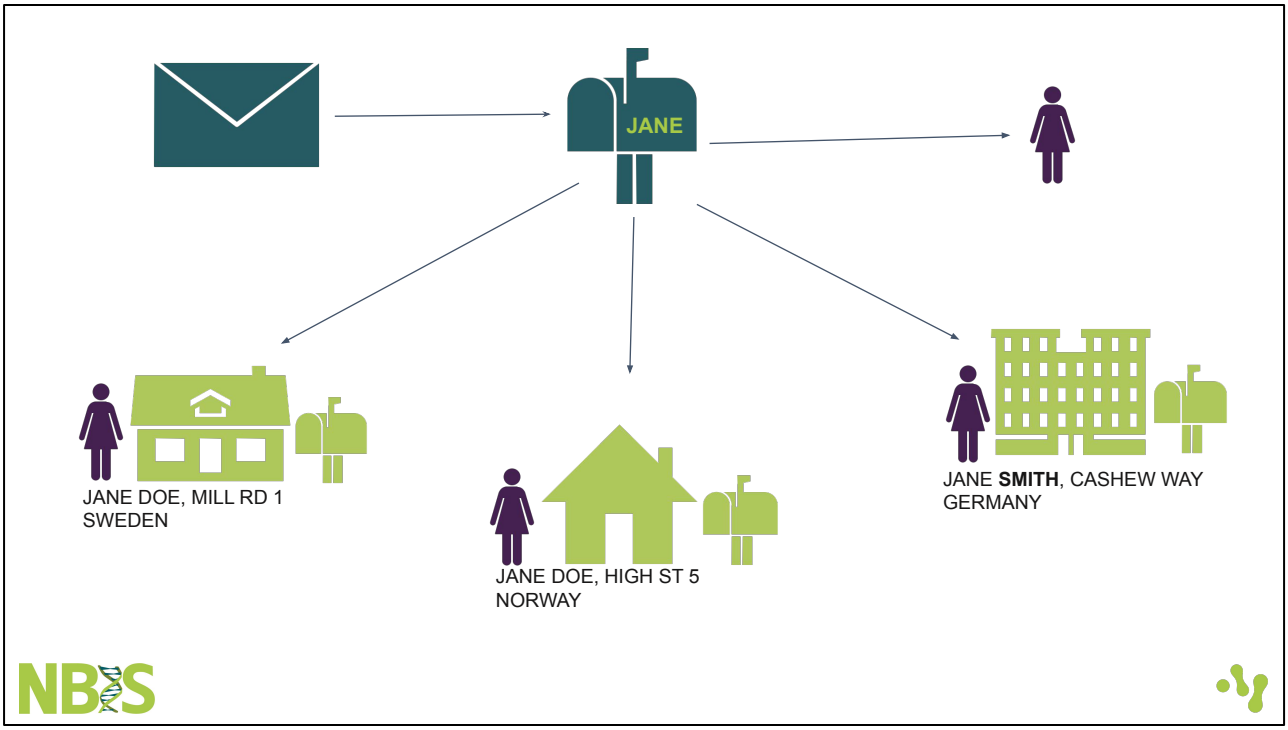
Identify

Verify

Locate

So to start with, what is a persistant identifier?
A persistent identifier (PID) is a long-lasting reference that uniquely tags a resource, for example a dataset, and provides the information required to reliably identify, verify and locate your resource eliminating many misunderstandings. (A PID may also be connected to a set of metadata which describes a digital or non-digital resource rather than to the item itself.) A PID ensures that access to the digital object is persistent. PIDs avoid broken links and difficulties locating a digital object (e.g. a dataset, a publication), even if its web address (URL) changes. A central registry ensures that following the PID will point you to the digital object's current location.

In order to make this a bit more concrete I'll give you an example.



Let’s imagine that you are going to be invited to a high school reunion in the hometown of your youth. However, you've moved several times in your life. With each move, your physical address changes. You might also have changed your name at some point in your life. When the organising committee wants to send you the invitation, they need your current name and address but since you both changed your name and moved your invite will be lost.

But what if there was a magical mailbox that stayed the same no matter where you moved? You could give out this "magical mailbox" address once, and anyone could send you letters to that address, which would then automatically find its way to wherever you live now. Even if you move across the globe, your mail reaches you through this magical, never-changing mailbox.

A Persistent Identifier (PID) acts like this magical mailbox for digital information. Let's say there's a specific research paper online. It gets a PID, a unique digital "address" that never changes, even if the paper moves to different locations on the internet (like from one database to another). No matter how the internet changes or where the document goes, you can use this PID to find the paper, just like how your letter finds you through your unchanging, magical mailbox.

The primary purpose of the PID is to provide the information required to reliably identify, verify and locate the resource it is connected with. In order to do so, the PIDs must comply with a few rules:

Features of PIDs

- Globally unique
  - It should comply with a **controlled syntax** to avoid clashes
- Persistent
  - It should be **maintained** for a **long period of time.** The syntax used for the identifier should also be persistent
- Resolvable
  - It should allow both **human** and **machine** users to **access the resource**

Globally unique:

To enable global uniqueness, a PID should comply with a **controlled syntax** to avoid clashes, for instance, by having **namespaces** that are **governed** by clearly defined **authorities**, ensuring that there are no two identical identifiers that point to different digital objects

Persistent:

The identifier, and the object to which it points, should be **maintained** for a **long period of time**. The syntax used for the identifier should be also persistent. A PID ensures that access to the digital object is persistent. PIDs avoid broken links and difficulties locating a digital object (e.g. a dataset, a publication), even if its web address (URL) changes. A central registry ensures that following the PID will point you to the digital object's current location.

Resolvable:

The identifier allows both **human** and **machine** users to **access the resource or information**

Digital Object Identifier - DOI

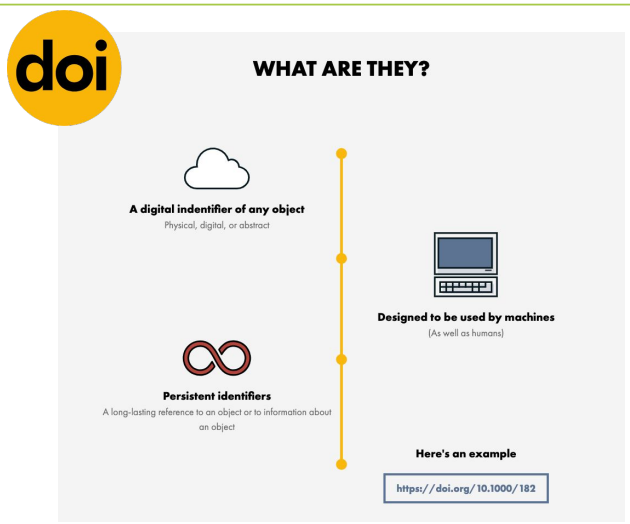


Illustration from the DOI Foundations website
https://www.doi.org/the-identifier/what-is-a-doi/

An example of a persistent identifier that you might be familiar with is the DOI **Digital Object Identifier (DOI)** (doi.org) It is one of the most common PIDs used by public repositories. A DOI, like any other PID, is a long-lasting reference that uniquely tags a resource. While the identifier itself is digital the object it is identifying can be of any kind, Physical, digital or abstract. DOIs are and are actionable, meaning that you can resolve it using the web browser and be taken to a web page with the listed digital object and its metadata. Actual access to the digital object from this page might be restricted since a PID may be connected to a set of metadata describing an item rather than to the item itself. DOIs are coupled with metadata that can be modified over time and to keep track of the locations and characteristics of the objects they identify.

Designed to be used by humans as well as machines, DOIs identify objects persistently. They allow things to be uniquely identified and accessed reliably. You know what you have, where it is, and others can track it too. A DOI is a unique number made up of a prefix and a suffix separated by a forward slash. This is an example of one: 10.1000/182. It is resolvable using our proxy server by displaying it as a link: https://doi.org/10.1000/182.

Foundation and RAs The DOI Foundation is the governance body of the DOI System and is composed of registration agencies that provide services to their respective communities (people or organisations who need to identify and track the things that matter to them.) Their work involves allocating DOI prefixes, registering DOI names, and providing a metadata schema associated with each DOI record. How does this work?

Digital Object Identifier - DOI

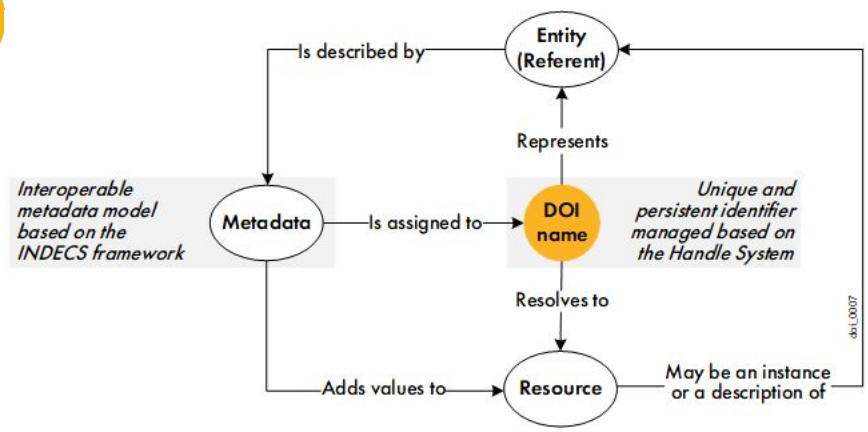


Illustration from The DOI Handbook April 2023 (https://doi.org/10.1000/182 identifies the latest current version of the handbook)

This figure illustrates the basic principle of the DOI System. Let's start in the middle. Any entity (digital, physical, or abstract) can be identified by a global unique and persistent identifier called a **DOI name**. The DOI name can be resolved to a resource, such as a web or internet resource for example a Zenodo record, with metadata describing the entity, a landing page with access to further resources, etc. The resource could also be an instance of the entity it represents. For example, a DOI name representing an article resolves to the web address of the HTML file version of the article.

Metadata must be assigned to each DOI name to describe the entity represented by the DOI name. Metadata interoperability is ensured through basic principles outlined by the DOI Foundation. Metadata is used to provide services to the users: it can be displayed to users to enrich an information resource; it can be used by users to search for a DOI name; etc.

Other relevant PIDs

ORCID - Open Researcher and Contributor ID

- persistent identifiers for researchers
- takes homonymy into account
- add aliases to your profile if your name changes
- ORCID stays the same when affiliation changes

The Research Organization Registry

- persistent identifiers for research organizations

ORCID, which stands for Open Researcher and Contributor ID, is a global, non-profit organisation which provides a **unique** and **persistent** identifier free of charge to researchers. Especially if you have a common name, you'll know how important it is to distinguish homonyms! It is extremely useful to be correctly identified, worldwide. ORCID takes homonymy into account, and the system also allows you to add aliases to your profile in the event that your name changes, making sure that it will be tracked back to you. Another benefit is that your ORCID will stay the same, even when your affiliation changes, ensuring that you get the credit you deserve and helping you keep track of your work.

ROR, The Research Organization Registry (ROR.org) is a global, community-led registry of open persistent identifiers for research organizations. ROR makes it easy for anyone or any system to disambiguate institution names and connect research organizations to researchers and research outputs. For example, Science for Life Laboratory is commonly referred to as SciLifeLab. If not directly involved with the organisation one might easily think that these are two different organisations, using the SciLifeLab ROR-id will make it clear they refer to the same organisation.

Benefits of PIDs

- uniquely distinguish resources from similar objects (F)
- a place to keep the metadata (F)
- machine actionable identifiers increase findability (F)
- resolves providing a way or information on how to access the object (A)
- enhances citability leading to easier reuse (R)

The benefits of using unique identifiers are several and by using a public repository to assign a DOI to your training material you benefit from efficient management and accurate tracking, as well as gaining the ability to more easily automate processes and collaborate with partners in your community. Furthermore, DOIs facilitate accurate citation and tracking of outputs and for individuals to get recognised for their works.

F:
- uniquely distinguished from similar objects
- a place to keep the metadata
- machine actionable identifiers increase findability

A: - resolves providing a way or information on how to access the object

I: - NA

R: - enhances citability leading to easier reuse

Reflection

In the context of training materials why are PIDs needed?

Which identifier should be used for each need?

PIDs can help distinguish between:

- different materials - **DOI**
- different versions of the same material - **DOI**
- different authors and contributors - **ORCID**
- different organisations - **ROR**

Discuss with neighbour or in plenum depending on time and accessibility needs

PIDs can help distinguish between:

- different materials - **DOI**
- different versions of the same material - **DOI**
- different authors and contributors - **ORCID**
- different organisations - **ROR**

Case Study

1. Group discussion 10 minutes
   a. Assign 1 person to take notes in the shared document
   b. Read through the use case assigned to your group from the FAIR training handbook
   c. Discuss and write down a short summary of the strategy used as well as pros and cons with this strategy

2. Plenary discussion 10 minutes
   a. Each group will share their observations and reflections

NBIS

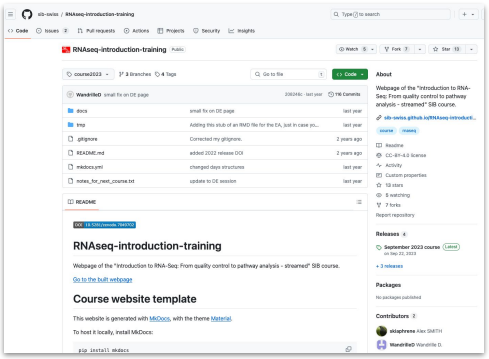
Execution:
- Divide students in 3 groups (or more if needed)
- Assign each group a use case
- Group discussions for 10 minutes
- Plenary discussion for 10 minutes (more time will be needed if you have more groups but if there are 2 groups with the same use case group 1 can summarize the strategy and group 2 can share their pros and cons, then group 1 can be invited to add any additional pros and cons they discussed)

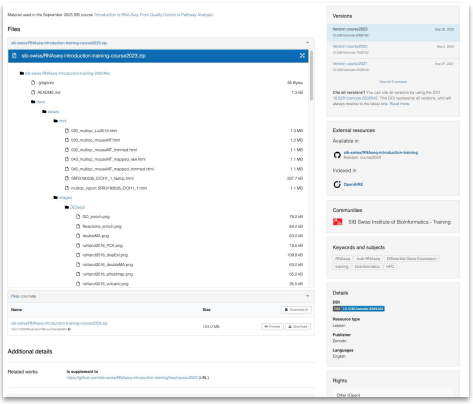
Instructions to students:
You will discuss one use case per group and then share this with the rest of the groups.
- Assign someone to take notes in the shared document
- Read the use case and discuss benefits and drawbacks of the strategy used in your use case
- Write down short summary of the strategy and the pros and

Preparations:
- In the collaborative notes document: Link to the use cases, sections for each group to write in

GitHub integrations

GitHub ⟷ zenodo





doi 10.5281/zenodo.5532542

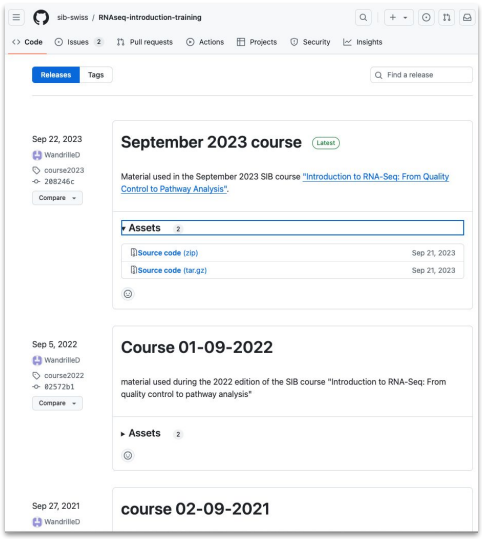
In the case studies you've seen examples of different strategies to utilize the public repository Zenodo to assign dois to training materials. That is one way of doing it.
In this course we have promoted the use of GitHub for hosting markdown based training material. The public repositories Zenodo and Figshare have integrations with GitHub set up in order to issue DOIs for repositories.

GitHub repository: https://github.com/sib-swiss/RNAseq-introduction-training

Zenodo Record: https://zenodo.org/doi/10.5281/zenodo.5532542

GitHub releases

- Snapshot of project at specific time point
- Packaged with re-use in mind
- Downloadable (zip file and tarball)
- Attached with version number/name via a tag
- Release notes to describe the specifics of the snapshot

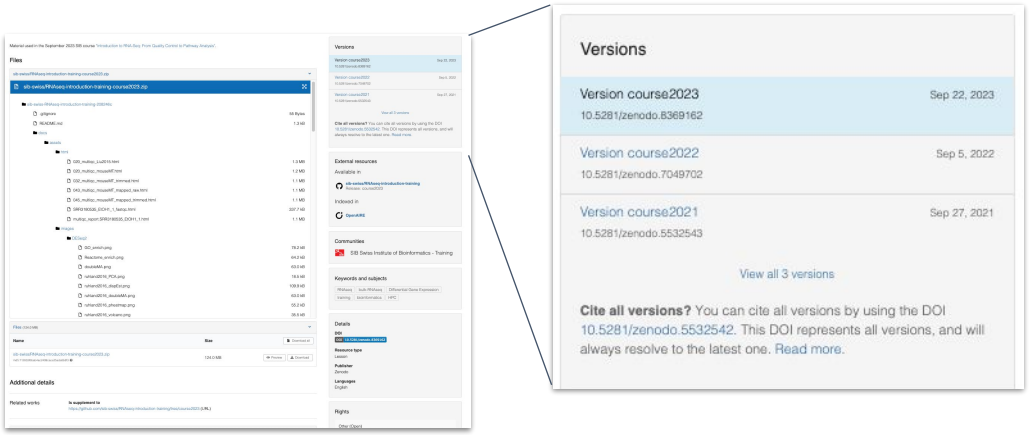


These integrations make use of GitHub releases which can be described as a snapshot of your project at a specific point in time that is packaged in way that make the content of the repository available for a wider audience to download and use. Each release is usually attached with a version number or name and attached to the snapshot in GitHub via a tag. This number/name helps users and authors keep track of different stages of the project and understand the differences between multiple releases. Releases can also be coupled with a title and release notes where the details of the specific release can be described. This will allow someone looking for a specific feature to find out what changes have been made without downloading and extracting the entire repository.

Since the integration between GitHub and Zenodo make use of releases, a new doi will be created for each new release of the connected repository. These dois are linked

GitHub documentation on releases:
https://docs.github.com/en/repositories/releasing-projects-on-github/about-releases
Github repository: https://github.com/sib-swiss/RNAseq-introduction-training/releases

Versioned DOIs with releases



Since the integration between GitHub and Zenodo make use of releases, a new DOI will be created for each new release of the connected repository. These DOIs are linked and there will be 1 DOI representing all versions that will always resolve to the latest one. Each version also has their own unique DOI.

Reflection

How would a good strategy for your own context look like?

Things to consider

- Do you want to get a PID for each training material?
- Do you want to get one PID for your whole training or course?
- Do you want to get a separate PID for each topic/module? For example, for a course containing several topics.
- Do you want to create a collection of topics with a PID where each concept will also have a PID and associated metadata?
- Do you want to get one PID for your whole training or course?

Individual reflection that might be left as an “after the course homework” depending on time.
If this is taught in a context where several ppl are from the same team or organisation it might be useful to do this as a group discussion.

Introduction to Tutorial

1. Use zenodo Sandbox to get a DOI for your training material
    a. By using GitHub integration
    b. By manually uploading a zip file of your GDrive folder
2. Enrich the Zenodo record with the metadata from previous session
3. Add the DOI to your hosting platform
4. Add the DOI to your TeSS record

Go to the tutorial in Chapter 08

https://elixir-europe-training.github.io/ELIXIR-TrP-FAIR-Material-By-Design/chapters/chapter_08/#84-tutorial-for-implementing-your-strategy